Stability-Constrained Markov Decision Processes Using MPC [⋆]

Mario Zanon [a], Sébastien Gros [b], Michele Palladino [c],

[a] *IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100, Lucca, Italy*

[b] *NTNU, Gløshaugen, Trondheim, Norway*

[c] *Gran Sasso Science Institute - GSSI, via Michele Jacobucci 2, 67100, L'Aquila, Italy*

Abstract

In this paper, we consider solving discounted Markov Decision Processes (MDPs) under the constraint that the resulting policy is stabilizing. In practice MDPs are solved based on some form of policy approximation. We will leverage recent results proposing to use Model Predictive Control (MPC) as a structured approximator in the context of Reinforcement Learning, which makes it possible to introduce stability requirements directly inside the MPC-based policy. This will restrict the solution of the MDP to stabilizing policies by construction. Because the stability theory for MPC is most mature for the undiscounted MPC case, we will first show in this paper that stable discounted MDPs can be reformulated as undiscounted ones. This observation will entail that the undiscounted MPC-based policy with stability guarantees will produce the optimal policy for the discounted MDP if it is stable, and the best stabilizing policy otherwise.

Key words: Markov Decision Processes, Model Predictive Control, Stability, Safe Reinforcement Learning

1 Introduction

Markov Decision Processes (MDPs) include a wide class of problems in which a controlled stochastic system needs to minimize a prescribed cost function (or maximize a reward). A special case is obtained for deterministic systems, in which case the problem is often labeled optimal control. MDPs have been extensively studied [6,7,23,25]; most of the existing literature focuses on studying the theoretical properties of MDPs from an optimization point of view, and on deriving algorithms to solve them, i.e., to compute optimal control policies.

Solving MDPs exactly is notoriously difficult, and practical approaches often rely on approximate Dynamic Programming or Reinforcement Learning (RL), using some form of function approximation [7,25]. The latter approximation approach has recently demonstrated the ability to solve problems that were previously considered intractable, see, e.g., [1,27]. A recently proposed function approximator for RL is Model Predictive Control (MPC) [13,14,16,28,29,32]. One of the benefits of using MPC as a function approximator is the existence of a strong theory proving desirable properties such as safety, stability and some form of explainability [17,24]. This fact has motivated recent interest in combining MPC with learning techniques, see, e.g., [3,4,10,18,21,26].

To the best of our knowledge, limited attention has been devoted to the enforcement of stability conditions in MDPs. The study of the stability properties of Markov Chains has been extensively studied in [20], while the stability properties of undiscounted optimal control for both deterministic systems and stochastic systems with bounded noise have been studied, e.g., in [17,24]. The derivation of conditions for the stability of discounted problems is harder and some results have been obtained for the deterministic case in, e.g., [12,22,31]. A promising new stability theory for generalized MDPs has been proposed in [15]. While this work considers

[⋆] This paper was not presented at any IFAC meeting. Corresponding author M. Zanon. This paper was partially supported by the Italian Ministry of University and Research under the PRIN'17 project "Data-driven learning of constrained control systems" , contract no. 2017J89ARP; by ARTES 4.0 Advanced Robotics and enabling digital Technologies & Systems 4.0, CUP: B81J18000090008; and by the Norwegian Research Council project "Safe Reinforcement-Learning using MPC" (SARLEM).

Email address: mario.zanon@imtlucca.it (Mario Zanon).

the undiscounted case, a combination with the results of [31] should make it possible to extend it to also cover the discounted case. The main drawback of the mentioned approaches is the extreme difficulty to use them in practice to guarantee stability in a rigorous way. Indeed, many of the proposed stability conditions are unfortunately very difficult to verify, which motivates the investigation of alternative approaches.

Given the aforementioned difficulties, in this paper we aim at imposing stability conditions by explicitly constraining the candidate policies to be stabilizing. We will do so by leveraging the existing stability theory for undiscounted MPC and the recent result of [13] which states that, under mild conditions, the MDP optimal policy can be captured via an MPC scheme with an approximate model provided that it is discounted by the same factor as the MDP. To the best of our knowledge, the direct introduction of stability constraints in an MDP has never been proposed before.

Main Contributions: Because we want to capture the solution of a discounted MDP using an undiscounted MPC, the theoretical gap between the discounted and undiscounted case must be closed first. To that end we will prove that, under a weak stability condition, a given discounted MDP can be formulated as an undiscounted MDP delivering the same optimal policies. En passant, we will also show that the corresponding undiscounted MDP corresponds to a bias optimal formulation, though this fact is not obvious at first sight. Using this result, the MPC-based policy, restricted to be stabilizing, will deliver the optimal policy for the discounted MDP if its optimal policies are stabilizing; and the optimal policy among the stabilizing policies of the discounted MDP if its optimal policies are not stabilizing.

This paper is structured as follows. In Section 2 we will provide the mathematical background and the required definitions. In Section 3 we will derive an undiscounted MDP whose optimal policy is also optimal for a given discounted MDP; we will relate the optimality criterion for this undiscounted MDP to commonly used optimality criteria; and we will illustrate the theory in the context of the linear quadratic regulator. We will then formulate the stability constraints in Section 4, where we will exploit the properties of MPC as a function approximator to solve MDPs. We will provide numerical examples in Section 5 and draw our conclusions in Section 6.

2 Preliminaries

We will consider that the problem is described by a Markov Decision Process (MDP) having the (possibly) stochastic state transition dynamics

$$\mathbb{P}\left[\boldsymbol{s}_+ \,|\, \boldsymbol{s},\boldsymbol{a}\right], \tag{1}$$

where $\boldsymbol{s},\boldsymbol{a}$ is the current state-action pair and $\boldsymbol{s}_+$ is the subsequent state. We will generally assume that the state-action space is continuous but the theory proposed here is valid in general. We observe that notation (1) is standard in the literature on MDPs, while the control literature typically uses the notation $\boldsymbol{s}_+ = \boldsymbol{f}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ is a stochastic variable and $\boldsymbol{f}$ a possibly nonlinear function. For discrete state spaces, (1) is a probability, while for continuous state spaces it is a probability density or measure.

We will label $L(\boldsymbol{s},\boldsymbol{a})$ the stage cost associated to the MDP, which we will assume can take the form

$$L\left(\boldsymbol{s},\boldsymbol{a}\right) = l\left(\boldsymbol{s},\boldsymbol{a}\right) + \mathcal{I}_\infty\left(\boldsymbol{h}\left(\boldsymbol{s},\boldsymbol{a}\right)\right), \tag{2}$$

where we use the indicator function

$$\mathcal{I}_\infty(\boldsymbol{x}) = \begin{cases} \infty & \text{if } \boldsymbol{x}_i > 0 \text{ for some } i \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

In (2), function l captures the cost given to different state-input pairs, while the constraints

$$\boldsymbol{h}(\boldsymbol{s},\boldsymbol{a}) \leq 0, \tag{4}$$

capture undesirable states and inputs, and infinite values are given to L when (4) is violated.

Assumption 1 *The cost $l(\boldsymbol{s},\boldsymbol{a})$ is finite for all finite states and inputs $\boldsymbol{s}$,$\boldsymbol{a}$. Additionally, for continuous state spaces,* (1) *satisfies*

$$\lim_{\alpha\to\infty} \mathbb{P}\left[\alpha \boldsymbol{s}_+ \,|\, \boldsymbol{s},\boldsymbol{a}\right] = 0, \qquad \forall\, \boldsymbol{s}_+,\boldsymbol{s},\boldsymbol{a} \textit{ finite.} \tag{5}$$

This mild assumption entails that the stage cost l remains bounded over finite horizons with probability 1. This allows us to avoid cumbersome technicalities in the proofs.

In this paper, we aim at solving the stability-constrained discounted MDP

$$\min_{\boldsymbol{\pi}\in\Pi_\mathrm{s}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right], \tag{6}$$

where Π_s denotes the set of all policies $\boldsymbol{\pi}$ which stabilize system (1). While our theory is valid for any stability definition, for the sake of simplicity in the illustrative examples we will propose in this paper, we will use the definition of asymptotic stability below. Note that the set Π_s then collects all policies $\boldsymbol{\pi}$ that yield asymptotic stability in the sense of Definition 1.

Definition 1 (Asymptotic Stability) *The set $\mathcal{A}$ is asymptotically stable if there exists a function $\beta \in \mathcal{KL}$ such that any state trajectory yielded by (1) in closed-loop with some policy $\boldsymbol{\pi}$, i.e., with $\boldsymbol{a} = \boldsymbol{\pi}(\boldsymbol{s})$, satisfies*

$$\|\boldsymbol{s}_k\|_{\mathcal{A}} \leq \beta\left(\|\boldsymbol{s}_0\|_{\mathcal{A}}, k\right),$$

where $\|\cdot\|_{\mathcal{A}}$ denotes the distance from set $\mathcal{A}$.

Notable specific cases are given by nominal asymptotic stability, in which set $\mathcal{A}$ is a single steady-state [24]; and robust asymptotic stability in which set $\mathcal{A}$ is the minimum robust positively invariant set [9,19,30]. Note that (robust) positive invariance of $\mathcal{A}$ is a necessary consequence of Definition 1.

Since in this paper we will compare discounted MDPs with undiscounted MDPs, we summarize next the corresponding optimality concepts. All definitions we will provide are given without stability constraints, but we remark that the same definitions hold also if the additional constraint $\boldsymbol{\pi} \in \Pi_{\mathrm{s}}$ is introduced. For more details on MDPs and optimality notions we refer to, e.g., [23] and references therein.

2.1 Optimality of Discounted MDPs

With the introduction of a discount factor $0 < \gamma \leq 1$, given (1) and (2), any optimal discounted policy $\boldsymbol{\pi}_\star^\gamma$ minimizes the expected total discounted cost

$$\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma := \arg\min_{\boldsymbol{\pi}} \; \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right], \quad (7)$$

where the expected value $\mathbb{E}_{\boldsymbol{\pi}}[\cdot]$ is taken over the state transition dynamics (1) in closed loop with policy $\boldsymbol{\pi}$. Since the policy need not be unique, $\boldsymbol{\Pi}_\star^\gamma$ is defined as a set. In the following, whenever we refer to a policy denoted as $\boldsymbol{\pi}_\star^\gamma$ we will implicitly assume that $\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma$. For any policy we define the associated action-value function $Q_{\boldsymbol{\pi}}^\gamma$ and value function $V_{\boldsymbol{\pi}}^\gamma$ as

$$\begin{aligned} V_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s}) &:= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k)) \,\middle|\, \boldsymbol{s}_0 = \boldsymbol{s}\right], && \text{(8a)} \\ Q_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s}, \boldsymbol{a}) &:= L(\boldsymbol{s}, \boldsymbol{a}) + \gamma \mathbb{E}\left[V_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s}_+) \,|\, \boldsymbol{s}, \boldsymbol{a}\right]. && \text{(8b)} \end{aligned}$$

The optimal action-value function $Q_\star^\gamma = Q_{\boldsymbol{\pi}_\star^\gamma}^\gamma$ and value function $V_\star^\gamma = V_{\boldsymbol{\pi}_\star^\gamma}^\gamma$ associated with the discounted MDP are defined by the Bellman equations [5]:

$$\begin{aligned} Q_\star^\gamma(\boldsymbol{s}, \boldsymbol{a}) &= L(\boldsymbol{s}, \boldsymbol{a}) + \gamma \mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+) \,|\, \boldsymbol{s}, \boldsymbol{a}\right], && \text{(9a)} \\ V_\star^\gamma(\boldsymbol{s}) &= Q_\star^\gamma(\boldsymbol{s}, \boldsymbol{\pi}_\star^\gamma(\boldsymbol{s})) = \min_{\boldsymbol{a}} Q_\star^\gamma(\boldsymbol{s}, \boldsymbol{a}). && \text{(9b)} \end{aligned}$$

Throughout the paper we will assume that the MDP underlying the system, the associated stage cost L and the discount factor γ yield a well-posed problem, i.e., the value functions defined by (9) are well-posed, and finite over some non-empty sets. This well-posedness is formulated in the following assumption.

Assumption 2 *There exists a nonempty set $\mathcal{S}$ such that for all $\boldsymbol{s} \in \mathcal{S}$ it holds that*

$$|V_\star^\gamma(\boldsymbol{s})| < \infty. \quad (10)$$

We state next an immediate consequence of this assumption, which will be useful afterwards.

Lemma 1 *Suppose that Assumptions 1-2 hold. Then,*

$$-\infty < \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[V_\star^\gamma(\boldsymbol{s}_k) \,|\, \boldsymbol{s}_0 = \boldsymbol{s}\right] < \infty, \quad \forall\, k = 0, \ldots, N, \quad (11)$$

holds for all $\boldsymbol{s} \in \mathcal{S}$, and N finite.

PROOF. We observe that

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\pi}_\star}\left[V_\star^\gamma(\boldsymbol{s}_k) \,|\, \boldsymbol{s}_0\right] = && \text{(12)} \\ &\quad \gamma^{-k} V_\star^\gamma(\boldsymbol{s}_0) - \gamma^{-k}\, \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[\sum_{i=0}^{k-1} \gamma^i L\left(\boldsymbol{s}_i, \boldsymbol{\pi}_\star^\gamma(\boldsymbol{s}_i)\right) \,\middle|\, \boldsymbol{s}_0\right]. \end{aligned}$$

Since $\gamma^{-k} V_\star^\gamma(\boldsymbol{s}_0)$ is bounded for any $k < \infty$ and $\boldsymbol{s}_0 \in \mathcal{S}$, due to Assumption 1 and Equation (5), the second term of the right-hand side of (12) must be lower bounded for all k finite. Additionally, in order for $V_\star^\gamma(\boldsymbol{s}_0)$ to be finite, it must also be upper bounded. Consequently, $\mathbb{E}_{\boldsymbol{\pi}_\star}\left[V_\star^\gamma(\boldsymbol{s}_k) \,|\, \boldsymbol{s}_0\right]$ must be bounded. □

2.2 Optimality of Undiscounted MDPs

For undiscounted MDPs, a discount factor $\gamma = 1$ is selected, such that the cost in (7) can be unbounded; and optimal policies are typically defined according to a hierarchy of criteria. The first criterion is based on the expected average total cost

$$\bar{\boldsymbol{\pi}}_\star \in \bar{\boldsymbol{\Pi}}_\star := \arg\min_{\boldsymbol{\pi}} \; \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N-1} L(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right]. \quad (13)$$

Since the policy need not be unique, $\bar{\boldsymbol{\Pi}}_\star$ is defined as a set. In the following, whenever we refer to a policy $\bar{\boldsymbol{\pi}}_\star$ we will implicitly assume that $\bar{\boldsymbol{\pi}}_\star \in \bar{\boldsymbol{\Pi}}_\star$. The *gain*, or average cost, is given by

$$\bar{L}_\infty(\boldsymbol{s}_0) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\bar{\boldsymbol{\pi}}_\star}\left[\sum_{k=0}^{N-1} L(\boldsymbol{s}_k, \bar{\boldsymbol{\pi}}_\star(\boldsymbol{s}_k)) \,\middle|\, \boldsymbol{s}_0\right]. \quad (14)$$

Any policy $\bar{\boldsymbol{\pi}}_\star$ satisfying (13) is called *gain optimal*. The average cost $\bar{L}_\infty$ is often assumed to be independent of the initial state $\boldsymbol{s}_0$ and we will also make this assumption for the sake of simplicity. For more information on the general case, we refer the interested reader to [23]. We ought to stress here that gain-optimal policies are in general not unique, such that more stringent optimality criteria might be introduced to further select among all gain-optimal policies.

The gain optimality criterion only ensures that the optimal average cost $\bar{L}_\infty$ is obtained, but it disregards transient performance. The concept of *bias optimality* has been introduced to account for the optimality of transients. A *bias optimal* policy $\tilde{\boldsymbol{\pi}}_\star$ minimizes the total undiscounted cost

$$\tilde{\boldsymbol{\pi}}_\star \in \tilde{\boldsymbol{\Pi}}_\star := \arg\min_{\boldsymbol{\pi}} \ \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty}\left(L(\boldsymbol{s}_k,\boldsymbol{\pi}(\boldsymbol{s}_k)) - \bar{L}_\infty\right)\right]. \tag{15}$$

Since also in this case the policy need not be unique, $\tilde{\boldsymbol{\Pi}}_\star$ is defined as a set. In the following, whenever we refer to a policy denoted as $\tilde{\boldsymbol{\pi}}_\star$ we will implicitly assume that $\tilde{\boldsymbol{\pi}}_\star \in \tilde{\boldsymbol{\Pi}}_\star$.

Due to the possible non-uniqueness, optimality notions which are more stringent than bias optimality have been introduced. These are beyond the scope of this paper; we simply recall that bias optimal policies are gain optimal and that the most stringent optimality criterion for undiscounted MDPs is *Blackwell optimality* [8,23].

3 Equivalent Undiscounted MDP Formulation

In this section, we discuss the equivalence between discounted MDPs and suitably formulated undiscounted MDPs. To that end, let us define the modified stage cost

$$\tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) := L(\boldsymbol{s},\boldsymbol{a}) + (\gamma-1)\mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{a}\right], \tag{16}$$

where we explicitly state the dependence on the discount factor γ to stress the fact that its definition is based on a discounted MDP formulation.

In the following, we construct the theory allowing one to support the discounted MDP solution by using undiscounted, finite-horizon stochastic MPC schemes. The latter are themselves undiscounted finite-horizon MDPs. Hence, it will be useful to first connect infinite-horizon discounted MDPs to finite-horizon undiscounted ones. We establish the equivalence between infinite-horizon discounted and infinite-horizon undiscounted MDPs later in the text.

In order to build these equivalences, let us define the N-steps undiscounted value and action-value functions as

$$\tilde{V}_{\boldsymbol{\pi}}^{\gamma,N}(\boldsymbol{s}) := \mathbb{E}_{\boldsymbol{\pi}}\left[V_\star^\gamma(\boldsymbol{s}_N) + \sum_{k=0}^{N-1}\tilde{L}^\gamma(\boldsymbol{s}_k,\boldsymbol{\pi}(\boldsymbol{s}_k))\,\middle|\,\boldsymbol{s}_0=\boldsymbol{s}\right], \tag{17a}$$

$$\tilde{Q}_{\boldsymbol{\pi}}^{\gamma,N}(\boldsymbol{s},\boldsymbol{a}) := \tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) + \mathbb{E}\left[\tilde{V}_{\boldsymbol{\pi}}^{\gamma,N-1}(\boldsymbol{s}_+)\,\middle|\,\boldsymbol{s},\boldsymbol{a}\right]. \tag{17b}$$

We define the corresponding (possibly non-unique) optimal policies as

$$\tilde{\boldsymbol{\pi}}_\star^{\gamma,N} \in \tilde{\boldsymbol{\Pi}}_\star^{\gamma,N} := \arg\min_{\boldsymbol{\pi}} \tilde{V}_{\boldsymbol{\pi}}^{\gamma,N}(\boldsymbol{s}), \tag{18}$$

with associated optimal value and action-value function $\tilde{V}_\star^{\gamma,N} := \tilde{V}^{\gamma,N}_{\tilde{\boldsymbol{\pi}}_\star^{\gamma,N}}$, and $\tilde{Q}_\star^{\gamma,N} := \tilde{Q}^{\gamma,N}_{\tilde{\boldsymbol{\pi}}_\star^{\gamma,N}}$. We will justify in Section 3.1 the use of the $\tilde{\cdot}$ notation, which we also used to define bias optimal quantities. We deliver next our first result, which proves the equivalence between the optimal undiscounted N-steps value functions and the optimal discounted value functions.

Theorem 1 *Suppose that Assumption 2 holds for all* $\boldsymbol{s} \in \mathcal{S}$*. Then* $\forall\, \boldsymbol{s} \in \mathcal{S}$, $\forall\, \boldsymbol{a}$, $N < \infty$ *it holds that*

$$\tilde{V}_\star^{\gamma,N}(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}), \qquad \tilde{Q}_\star^{\gamma,N}(\boldsymbol{s},\boldsymbol{a}) = Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}).$$

PROOF. For any $\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma$, we use the Bellman Equation (9) to obtain

$$L(\boldsymbol{s},\boldsymbol{\pi}_\star^\gamma(\boldsymbol{s})) = V_\star^\gamma(\boldsymbol{s}) - \gamma\mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{\pi}_\star^\gamma(\boldsymbol{s})\right], \tag{19}$$

which we use together with (16) to obtain

$$\tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{\pi}_\star^\gamma(\boldsymbol{s})) = V_\star^\gamma(\boldsymbol{s}) - \mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{a}\right]. \tag{20}$$

Equation (17a) then becomes the telescopic sum:

$$\tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^{\gamma,N}(\boldsymbol{s}_0) = \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[V_\star^\gamma(\boldsymbol{s}_N) + \sum_{k=0}^{N-1} V_\star^\gamma(\boldsymbol{s}_k) - V_\star^\gamma(\boldsymbol{s}_{k+1})\right].$$

Using Assumption 2, we simplify the terms in the telescopic sum to obtain

$$\tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^{\gamma,\bar{N}}(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}), \qquad \forall\, \bar{N} \leq N. \tag{21}$$

Consequently, by (17b) we have

$$\begin{aligned}\tilde{Q}_{\boldsymbol{\pi}_\star^\gamma}^{\gamma,N}(\boldsymbol{s},\boldsymbol{a}) &= \tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) + \mathbb{E}\left[\tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^{\gamma,N-1}(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{a}\right]\\ &= L(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{a}\right] = Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}),\end{aligned}$$

where we exploited (16) and (21). Since

$$\boldsymbol{\Pi}_\star^\gamma = \arg\min_{\boldsymbol{a}} Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) = \arg\min_{\boldsymbol{a}} \tilde{Q}_{\boldsymbol{\pi}_\star^\gamma}^{\gamma,N}(\boldsymbol{s},\boldsymbol{a}) = \tilde{\boldsymbol{\Pi}}_\star^{\gamma,N},$$

we immediately obtain

$$\begin{aligned}\tilde{V}_\star^{\gamma,N}(\boldsymbol{s}) &= \tilde{V}^{\gamma,N}_{\tilde{\boldsymbol{\pi}}_\star^{\gamma,N}}(\boldsymbol{s}) = \tilde{V}^{\gamma,N}_{\boldsymbol{\pi}_\star^\gamma}(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}),\\ \tilde{Q}_\star^{\gamma,N}(\boldsymbol{s},\boldsymbol{a}) &= \tilde{Q}^{\gamma,N}_{\tilde{\boldsymbol{\pi}}_\star^{\gamma,N}}(\boldsymbol{s},\boldsymbol{a}) = \tilde{Q}^{\gamma,N}_{\boldsymbol{\pi}_\star^\gamma}(\boldsymbol{s},\boldsymbol{a}) = Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}).\end{aligned}$$

□

If the value function $\tilde{V}_\star^{\gamma,N}$ remains bounded for all $\boldsymbol{s} \in \mathcal{S}$ as $N \to \infty$, then the result above holds also as $N \to \infty$, but the terminal cost $V_\star^\gamma$ is still required in forming $\tilde{V}_{\boldsymbol{\pi}}^{\gamma,N}$, $\tilde{Q}_{\boldsymbol{\pi}}^{\gamma,N}$. In order to dismiss that terminal cost for $N \to \infty$ we need an additional stronger assumption.

Assumption 3 *For any $\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma$, there exists a nonempty set $\mathcal{S}$ such that for all $\boldsymbol{s} \in \mathcal{S}$ it holds that*

$$-\infty < \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[V_\star^\gamma(\boldsymbol{s}_k) \,|\, \boldsymbol{s}_0 = \boldsymbol{s}\right] < \infty, \quad \forall\, k = 0, \dots, \infty. \tag{22}$$

Moreover,

$$-\infty < \lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[V_\star^\gamma(\boldsymbol{s}_k) \,|\, \boldsymbol{s}_0\right] := v_\infty^\gamma < \infty, \tag{23}$$

holds $\forall\, \boldsymbol{s}_0 \in \mathcal{S}$ for some constant v_∞^γ.

Assumption 3 entails a weak form of stability for the discounted MDP, as we discuss next. It is stronger than requiring the existence of a bounded value function, i.e.,

$$\left|\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}\left[\sum_{k=0}^{N-1} \gamma^k L(\boldsymbol{s}_k, \boldsymbol{\pi}_\star^\gamma(\boldsymbol{s}_k)) \,\middle|\, \boldsymbol{s}_0\right]\right| \leq \infty, \tag{24}$$

holds for any $\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma$ on a non-empty set of initial conditions $\boldsymbol{s}_0$. Indeed, (24) is finite provided that the stage cost $L(\boldsymbol{s}_k, \boldsymbol{\pi}_\star^\gamma(\boldsymbol{s}_k))$ diverges in expectation at a rate no larger than γ_{D}^{-k} for some $\gamma_{\mathrm{D}} > \gamma$. It follows that (24) allows the cost to grow unbounded over time. Hence, the existence of a bounded value function (Assumption 2) does not entail that the value function remains bounded over an infinite time. Moreover, assuming (23), i.e., that the limit exists and converges to the constant v_∞^γ, introduces further restrictions, which rule out, e.g., periodic oscillations of the cost. We cast Assumption 3 as a weak form of stability: if the MDP converges with finite cost to a unique steady-state distribution, then Assumption 3 automatically holds. However, converging to a steady-state distribution is a stronger requirement, since Assumption 3 might hold also for non-steady-state and even diverging distributions, see, e.g., Section 3.2.

In order to formulate the next theorem, let us use (16) to first define the undiscounted value and action-value functions without terminal cost as

$$\tilde{V}_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s}) := \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N-1} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k)) \,\middle|\, \boldsymbol{s}_0 = \boldsymbol{s}\right], \tag{25a}$$

$$\tilde{Q}_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s},\boldsymbol{a}) := \tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) + \mathbb{E}\left[\tilde{V}_{\boldsymbol{\pi}}^\gamma(\boldsymbol{s}_+) \,\middle|\, \boldsymbol{s},\boldsymbol{a}\right], \tag{25b}$$

which do not necessarily match the limit for $N \to \infty$ of the value functions defined in (17). Furthermore, we define the optimal undiscounted policies and the associated value functions as:

$$\tilde{\boldsymbol{\pi}}_\star^\gamma \in \tilde{\boldsymbol{\Pi}}_\star^\gamma := \arg\min_{\boldsymbol{\pi}} \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N-1} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right], \tag{26a}$$

$$\tilde{V}_\star^\gamma(\boldsymbol{s}) := \tilde{V}_{\tilde{\boldsymbol{\pi}}_\star^\gamma}^\gamma(\boldsymbol{s}), \qquad \tilde{Q}_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) := \tilde{Q}_{\tilde{\boldsymbol{\pi}}_\star^\gamma}^\gamma(\boldsymbol{s},\boldsymbol{a}). \tag{26b}$$

Theorem 2 *Suppose that Assumption 3 holds. Then $\forall\, \boldsymbol{s} \in \mathcal{S}$, $\forall\, \boldsymbol{a}$ it holds that*

$$\tilde{V}_\star^\gamma(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}) - v_\infty^\gamma, \quad \tilde{Q}_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) = Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) - v_\infty^\gamma, \tag{27a}$$

$$\tilde{\boldsymbol{\Pi}}_\star^\gamma(\boldsymbol{s}) = \boldsymbol{\Pi}_\star^\gamma(\boldsymbol{s}). \tag{27b}$$

PROOF. Similar to the proof of Theorem 1, we use the Bellman Equation (9) together with (19)-(20) to write (25a) as a telescopic sum in which all terms are bounded due to Assumption 3. By simplifying the terms in the sum we obtain that, for any $\boldsymbol{\pi}_\star^\gamma \in \boldsymbol{\Pi}_\star^\gamma$,

$$\tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}) - \lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star^\gamma}[V_\star^\gamma(\boldsymbol{s}_k)] = V_\star^\gamma(\boldsymbol{s}) - v_\infty^\gamma. \tag{28}$$

Consequently,

$$\begin{aligned}\tilde{Q}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s},\boldsymbol{a}) &= \tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) + \mathbb{E}\left[\tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s}_+) | \boldsymbol{s},\boldsymbol{a}\right]\\ &= L(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+) | \boldsymbol{s},\boldsymbol{a}\right] - v_\infty^\gamma\\ &= Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) - v_\infty^\gamma,\end{aligned}$$

where we used (16) and (28). Then,

$$\tilde{\boldsymbol{\Pi}}_\star^\gamma = \arg\min_{\boldsymbol{a}} \tilde{Q}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s},\boldsymbol{a}) = \arg\min_{\boldsymbol{a}} Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) = \boldsymbol{\Pi}_\star^\gamma, \tag{29}$$

which immediately entails (27b) and, in turn,

$$\begin{aligned}\tilde{V}_\star^\gamma(\boldsymbol{s}) &= \tilde{V}_{\tilde{\boldsymbol{\pi}}_\star^\gamma}^\gamma(\boldsymbol{s}) = \tilde{V}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s}) = V_\star^\gamma(\boldsymbol{s}) - v_\infty^\gamma,\\ \tilde{Q}_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) &= \tilde{Q}_{\tilde{\boldsymbol{\pi}}_\star^\gamma}^\gamma(\boldsymbol{s},\boldsymbol{a}) = \tilde{Q}_{\boldsymbol{\pi}_\star^\gamma}^\gamma(\boldsymbol{s},\boldsymbol{a}) = Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) - v_\infty^\gamma.\end{aligned}$$

□

This theorem establishes that any discounted MDP satisfying Assumption 3 can be reformulated as an undiscounted MDP which delivers the same policy and the same value functions up to a constant term.

We ought to stress here that in the following we are interested in forming policies which are stabilizing by construction, i.e., we aim at solving the discounted MDP under the constraint of preserving stability. The equivalence proposed in Theorem 2 will be instrumental in allowing us to formulate such constraints in the undiscounted setting, while optimizing the cost in a discounted sense. This is of particular interest, since the stability analysis is much simpler and more developed for the undiscounted setting. We will discuss the introduction of stability constraints in Section 4. Before detailing how one can introduce stability constraints, we first prove that the obtained undiscounted MDP (25)-(26) yields bias optimal policies; and we then illustrate the theoretical developments in the simple case of a Linear Quadratic Regulator (LQR).

3.1 Equivalence of Optimality Notions

The undiscounted MDP used in Theorem 2 minimizes the cost in (26a), which is not directly related to standard optimality concepts such as gain or bias optimality. We therefore prove next that the policy $\tilde{\boldsymbol{\pi}}^\gamma_\star = \boldsymbol{\pi}^\gamma_\star$ obtained from (26a) is in fact bias optimal. To that end, we will first prove gain optimality. We will then prove that the optimal gain is 0, which we will relate to bias optimality.

Theorem 3 *Suppose that Assumption 3 holds. Then, any $\boldsymbol{\pi}^\gamma_\star \in \boldsymbol{\Pi}^\gamma_\star$ is gain optimal for stage cost $\tilde{L}^\gamma$.*

PROOF. By Theorem 2 policy $\boldsymbol{\pi}^\gamma_\star$ solves (26a) with a finite optimal cost, such that $\forall\, \boldsymbol{\pi}$ we have

$$\mathbb{E}_{\boldsymbol{\pi}^\gamma_\star}\left[\sum_{k=0}^{\infty} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right] \leq \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right]. \quad (30)$$

Therefore, the policy is gain optimal, i.e., $\forall\, \boldsymbol{\pi}$ we have

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\boldsymbol{\pi}^\gamma_\star}&\left[\sum_{k=0}^{N-1} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right] \\ &\leq \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N-1} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right].\end{aligned}$$

□

Theorem 4 *Suppose that Assumption 3 holds. Then, any $\boldsymbol{\pi}^\gamma_\star \in \boldsymbol{\Pi}^\gamma_\star$ is bias optimal for stage cost $\tilde{L}^\gamma$.*

PROOF. Because $\tilde{V}^\gamma_\star(\boldsymbol{s}_0)$ is finite for all $\boldsymbol{s}_0 \in \mathcal{S}$, we have that the average cost is

$$\begin{aligned}\bar{\tilde{L}}^\gamma_\infty &= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{k=0}^{N-1} \tilde{L}^\gamma(\boldsymbol{s}_k, \tilde{\boldsymbol{\pi}}^\gamma_\star(\boldsymbol{s}_k))\right] \\ &= \lim_{N\to\infty} \frac{1}{N}\tilde{V}^\gamma_\star(\boldsymbol{s}_0) = 0. \end{aligned} \quad (31)$$

By Theorem 3, any policy $\boldsymbol{\pi}^\gamma_\star \in \boldsymbol{\Pi}^\gamma_\star = \tilde{\boldsymbol{\Pi}}^\gamma_\star$ is gain optimal, such that by (31) the gain-optimal average cost is $\bar{\tilde{L}}^\gamma_\infty = 0$. Therefore, bias optimality for $\tilde{L}^\gamma$ is obtained by minimizing

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k)) - \bar{\tilde{L}}^\gamma_\infty\right] = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \tilde{L}^\gamma(\boldsymbol{s}_k, \boldsymbol{\pi}(\boldsymbol{s}_k))\right], \quad (32)$$

which is the total undiscounted reward. Since by Theorem 2 any policy $\boldsymbol{\pi}^\gamma_\star \in \boldsymbol{\Pi}^\gamma_\star = \tilde{\boldsymbol{\Pi}}^\gamma_\star$ solves (26a), i.e., minimizes (32), such that it is bias optimal by construction.

□

3.2 The LQR Case

In order to clarify the previous developments, let us consider the case of a stochastic linear system

$$\boldsymbol{s}_+ = A\boldsymbol{s} + B\boldsymbol{a} + \boldsymbol{w}, \quad (33)$$

with $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, W)$ i.i.d., $\mathbb{E}\left[\boldsymbol{w}\boldsymbol{s}^\top\right] = 0$, $\mathbb{E}\left[\boldsymbol{w}\boldsymbol{a}^\top\right] = 0$,

$$L(\boldsymbol{s}, \boldsymbol{a}) = \begin{bmatrix}\boldsymbol{s} \\ \boldsymbol{a}\end{bmatrix}^\top H \begin{bmatrix}\boldsymbol{s} \\ \boldsymbol{a}\end{bmatrix}, \qquad H = \begin{bmatrix} T & U^\top \\ U & R\end{bmatrix} \succ 0.$$

The value and action-value function are given by

$$\begin{aligned} V^\gamma_\star(\boldsymbol{s}) &= \boldsymbol{s}^\top P \boldsymbol{s} + V_0, \qquad V_0 = \frac{\gamma}{1-\gamma}\mathrm{Tr}\left(PW\right), \\ Q^\gamma_\star(\boldsymbol{s}, \boldsymbol{a}) &= \begin{bmatrix}\boldsymbol{s} \\ \boldsymbol{a}\end{bmatrix}^\top \begin{bmatrix} T + \gamma A^\top P A & U^\top + \gamma A^\top P B \\ U + \gamma B^\top P A & R + \gamma B^\top P B\end{bmatrix} \begin{bmatrix}\boldsymbol{s} \\ \boldsymbol{a}\end{bmatrix} + V_0, \end{aligned}$$

with optimal policy $\boldsymbol{\pi}^\gamma_\star(\boldsymbol{s}) = -K\boldsymbol{s}$, and

$$P = T + \gamma A^\top P A - (U^\top + \gamma A^\top P B)K, \quad (34a)$$

$$K = (R + \gamma B^\top P B)^{-1}(U + \gamma B^\top P A), \quad (34b)$$

$$0 \prec R + \gamma B^\top P B. \quad (34c)$$

Note that all elements of P can be finite even in case $\rho(A - BK) \geq 1$, where $\rho(\cdot)$ denotes the spectral radius, in which case the value function is defined and bounded for bounded states, but if P is full rank, Assumption 3 does not hold.

Checking Assumption 3 We observe that, under feedback $\boldsymbol{a} = -K\boldsymbol{s}$ we have

$$S_+ = A_K S A_K^\top + W,$$

where we used $A_K := A - BK$, $S := \mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}^\top\right]$. If $\rho(A_K) < 1$, then there exists a unique matrix S_∞ solving the Lyapunov equation $A_K S_\infty A_K^\top - S_\infty + W = 0$. Assume that P is full rank, then we have

$$\lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star}[V_\star^\gamma(\boldsymbol{s}_k)] = \begin{cases} \infty & \text{if } \rho(A_K) \geq 1 \\ \mathrm{Tr}\left(PS_\infty\right) + V_0 < \infty & \text{otherwise} \end{cases}.$$

The condition $\rho(A_K) < 1$ distinguishes the case $\lim_{k\to\infty} \mathbb{E}\left[\boldsymbol{s}_k\right] = 0$, $\lim_{k\to\infty} \mathbb{E}\left[\boldsymbol{s}_k\boldsymbol{s}_k^\top\right] = S_\infty$ from the case $\lim_{k\to\infty} \mathbb{E}\left[\boldsymbol{s}_k\right] = \pm\infty$: in the former, convergence in expectation of the value function is guaranteed; in the latter we immediately have $\lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star}\left[V_\star^\gamma(\boldsymbol{s}_k)\right] = \infty$.

Undiscounted Equivalent MDP We observe that

$$\begin{aligned}&\mathbb{E}\left[V_\star^\gamma(\boldsymbol{s}_+)|\boldsymbol{s},\boldsymbol{a}\right] \\ &\quad= \begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{a}\end{bmatrix}^\top \begin{bmatrix} A^\top PA & A^\top PB \\ B^\top PA & B^\top PB\end{bmatrix} \begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{a}\end{bmatrix} + \frac{1}{1-\gamma}\mathrm{Tr}\left(PW\right).\end{aligned}$$

In case Assumption 3 holds, the stage cost for the corresponding undiscounted MDP is then given by

$$\tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) = \begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{a}\end{bmatrix}^\top \tilde{H} \begin{bmatrix}\boldsymbol{s}\\ \boldsymbol{a}\end{bmatrix} - \mathrm{Tr}\left(PW\right), \tag{35}$$

$$\tilde{H} = \begin{bmatrix} T + (\gamma-1)A^\top PA & U^\top + (\gamma-1)A^\top PB \\ U + (\gamma-1)B^\top PA & R + (\gamma-1)B^\top PB\end{bmatrix}.$$

Consequently, we have

$$\begin{aligned}\tilde{V}_\star^\gamma(\boldsymbol{s}) &= V_\star^\gamma(\boldsymbol{s}) - \mathrm{Tr}\left(PS_\infty\right) - V_0 = \boldsymbol{s}^\top P\boldsymbol{s} - \mathrm{Tr}\left(PS_\infty\right), \\ \tilde{Q}_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) &= Q_\star^\gamma(\boldsymbol{s},\boldsymbol{a}) - \mathrm{Tr}\left(PS_\infty\right) - V_0.\end{aligned}$$

Then, if P is full rank, the following holds:

$$\lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\tilde{V}_\star^\gamma(\boldsymbol{s}_k)\right] = \begin{cases} \infty & \text{if } \rho(A-BK) \geq 1 \\ 0 & \text{otherwise}\end{cases}.$$

In case $P = 0$, Assumption 3 is automatically satisfied, even though the Markov chain might diverge. A simple example is given by $A = 2$, $B = 1$, $T = 0$, $U = 0$, $R = 1$, for any $\gamma \in]0,1]$. The system is unstable since $K = 0$ implies $A - BK = 2$, but $V_\star^\gamma(\boldsymbol{s}) = 0$ such that $\lim_{k\to\infty} V_\star^\gamma(\boldsymbol{s}) = 0$ holds. In this case we have $\tilde{L}^\gamma(\boldsymbol{s},\boldsymbol{a}) = L(\boldsymbol{s},\boldsymbol{a})$.

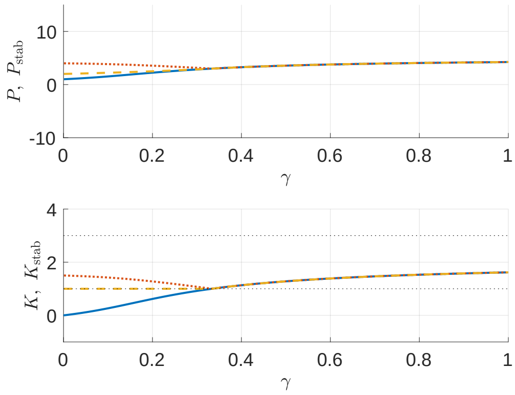

Fig. 1. Analytic solution of the LQR problem: unconstrained solution (blue continuous line), stability constrained solution (yellow dashed line), solution with $\tilde{L}$ and no terminal cost (dotted red line), and stability bounds (black dotted lines).

A Simple Example Consider a linear system of the form (33) with $A = 2$, $B = 1$, $T = 1$, $U = 0$, $R = 1$, and set $W = 0$ in order to have a deterministic system. One can verify that this system is stabilized only for feedback matrices $K \in \mathcal{K} :=]1,3[$, since this implies $A - BK \in]-1,1[$. The discounted LQR solution is

$$P = \frac{5\gamma - 1 + \sqrt{(1-5\gamma)^2 + 4\gamma}}{2\gamma}, \tag{36a}$$

$$K = \frac{4\gamma}{1 - 3\gamma + \sqrt{(1-5\gamma)^2 + 4\gamma}}, \tag{36b}$$

such that $\gamma \in]1/3, 1] \implies K \in \mathcal{K}$.

We can write the stability-constrained MDP problem as

$$K_{\rm stab} := \arg\min_K \ P_K(K) \qquad \text{s.t. } K \in \mathcal{K}, \tag{37}$$

where $V_{-K\boldsymbol{s}}^\gamma(\boldsymbol{s}) = P_K(K)\boldsymbol{s}^2$ and $P_K(K) := \frac{K^2+1}{1-\gamma(2-K)^2}$. We ought to stress that (37) is not well-posed, since its constraint set is open and a solution might not exist. In practice, one usually defines a closed subset of the open set $\mathcal{K}$ in order to (a) make the problem well-posed; and (b) avoid being too close to the stability margin and introduce some robustness to numerical errors. Since the closure $\mathrm{cl}\left(\mathcal{K}\right) := [1,3]$ of $\mathcal{K}$ guarantees marginal stability of the closed-loop system, we discuss the solution of

$$K_{\rm mstab} := \arg\min_K \ P_K(K) \qquad \text{s.t. } K \in \mathrm{cl}\left(\mathcal{K}\right).$$

One can verify that, using K given by (36b),

$$K_{\rm mstab} = \begin{cases} 1 & \gamma \leq 1/3 \\ K & \gamma > 1/3\end{cases}.$$

The solution is shown in Figure 1. The undiscounted equivalent problem is obtained by computing stage cost $\tilde{L}^\gamma$ as per (35). Note that, even though $\tilde{L}^\gamma$ is defined for all γ provided that $V_\star^\gamma$ is bounded, Theorem 2 applies only if Assumption 3 holds, which is not the case for $\gamma < 1/3$. Indeed, if $\gamma < 1/3$ the closed-loop system becomes unstable and $\lim_{k\to\infty} V_\star^\gamma(\boldsymbol{s}_k) = \infty$. In turn, this entails that in such cases the stability-constrained discounted MDP with stage cost L does not have the same solution as the stability-constrained undiscounted MDP with stage cost $\tilde{L}^\gamma$. This fact can be observed in Figure 1. We stress that if we formulate the undiscounted MDP with stage cost $\tilde{L}^\gamma$ and terminal cost $V_\star^\gamma(\boldsymbol{s})$, then Theorem 1 applies and the equivalence between the two MDPs holds for all γ. This situation is captured by the discrete algebraic Riccati equation, which has two solutions in this case, corresponding to the two MDP solutions: with and without the terminal cost $V_\star^\gamma$.

Finally, we ought to stress that, as well-known, if we consider the same system with nonzero covariance $W \neq 0$, the optimal feedback coincides with the one of its deterministic counterpart, i.e., when $W = 0$.

4 Stability Constraints based on MPC

In the previous section, we proved that a discounted MDP can be reformulated as an undiscounted MDP and we mentioned that this fact can be useful to introduce stability constraints in the MDP formulation. However, we did not discuss how this can be done in the general case. In this section we exploit our new theoretical results to propose a solution method for stability-constrained MDPs which can be implemented in practice.

In order to solve an MDP, one must compute either the value function $V_\star^\gamma$, the policy $\boldsymbol{\pi}_\star^\gamma$, the action-value function $Q_\star^\gamma$ or a combination of these. Since their functional form is not known a priori and can be rather complicated, solving MDPs exactly is a notoriously difficult task. Therefore, practical approaches typically rely on some parametric function approximation $V_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, and $Q_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ denotes a set of parameter adjusting the function approximations, in order to make the problem tractable [6,7]. The stability-constrained discounted MDP (6) can then be formulated using function approximation as

$$\min_{\boldsymbol{\theta}\in\Theta_\mathrm{s}} \quad \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\sum_{k=0}^{\infty}\gamma^k L(\boldsymbol{s}_k, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}_k))\right], \tag{38}$$

where $\Theta_\mathrm{s} := \{\ \boldsymbol{\theta} \mid \boldsymbol{\pi}_{\boldsymbol{\theta}} \in \Pi_\mathrm{s}\ \}$ is the set of parameters which yield a stabilizing policy.

We propose to introduce the stability constraint $\boldsymbol{\theta} \in \Theta_\mathrm{s}$ by relying on Model Predictive Control (MPC) to support the necessary function approximations. This approach has been proposed first in [13] in the context of reinforcement learning. MPC provides a very convenient way to support a parametric approximation of $V_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, and $Q_{\boldsymbol{\theta}}$, as it solves the optimal control problem

$$\begin{aligned}
Q_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{a}) = \min_{\boldsymbol{z}}\ & \lambda_{\boldsymbol{\theta}}(\boldsymbol{s}) + V_{\boldsymbol{\theta}}^\mathrm{f}(\boldsymbol{x}_N) + \sum_{k=0}^{N-1}\ell_{\boldsymbol{\theta}}(\boldsymbol{x}_k,\boldsymbol{u}_k) && \text{(39a)}\\
\text{s.t. } & \boldsymbol{x}_0 = \boldsymbol{s}, \quad \boldsymbol{u}_0 = \boldsymbol{a}, && \text{(39b)}\\
& \boldsymbol{x}_{k+1} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_k,\boldsymbol{u}_k), \quad k\in\mathbb{I}_0^{N-1}, && \text{(39c)}\\
& \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_k,\boldsymbol{u}_k) \leq 0, \qquad k\in\mathbb{I}_0^{N-1}, && \text{(39d)}\\
& \boldsymbol{h}_{\boldsymbol{\theta}}^\mathrm{f}(\boldsymbol{x}_N) \leq 0, && \text{(39e)}
\end{aligned}$$

where $\boldsymbol{z} = (\boldsymbol{x}_0, \boldsymbol{u}_0, \ldots, \boldsymbol{x}_N)$ and $\mathbb{I}_a^b$ denotes the set of integers larger than a and smaller than b. The stage and terminal cost $\ell_{\boldsymbol{\theta}}$, $V_{\boldsymbol{\theta}}^\mathrm{f}$, the system dynamics and constraints $\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{h}_{\boldsymbol{\theta}}, \boldsymbol{h}_{\boldsymbol{\theta}}^\mathrm{f}$ and the initial cost $\lambda_{\boldsymbol{\theta}}$ are all parametric functions of $\boldsymbol{\theta}$. Note that in MPC the initial constraint (39b) typically only involves the state, i.e., $\boldsymbol{u}_0 = \boldsymbol{a}$ is not present, since the goal is to compute an optimal policy. The policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s})$ and value function $V_{\boldsymbol{\theta}}(\boldsymbol{s})$ are obtained by solving Problem (39) with constraint $\boldsymbol{u}_0 = \boldsymbol{a}$ removed. This is equivalent to

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}) = \arg\min_{\boldsymbol{a}} Q_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{a}), \quad V_{\boldsymbol{\theta}}(\boldsymbol{s}) = \min_{\boldsymbol{a}} Q_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{a}).$$

We briefly comment on this particular MPC formulation. Function $\lambda_{\boldsymbol{\theta}}$ has been introduced in [13] to make it possible to use a positive-definite stage cost $\ell_{\boldsymbol{\theta}}$ even when the true stage cost l is not. This choice is related to the stability theory of economic MPC, where λ is called a *storage function*. In [13] it is discussed in detail how the use of a parameterized stage cost $\ell_{\boldsymbol{\theta}}$ makes it possible to recover the optimal policy and value functions using (36)-(37) even if the MPC model (39c) does not accurately capture the system dynamics (1). Finally, actuator limitations, though included in $\boldsymbol{h}_{\boldsymbol{\theta}}$ for ease of notation, are assumed to be known a priori and fixed, i.e., independent of $\boldsymbol{\theta}$. Note that this assumption is ubiquitous in RL and the possibility to also learn actuator limitations is often avoided as it poses several difficulties.

The specific formulation of set Θ_s depends on the specific MPC formulation used and the selected stability concept. We will provide some examples in Section 5.

We ought to stress that, though we formulated (39) using a notation which is easily interpreted as a deterministic formulation, any MPC formulation can be used, including stochastic and robust formulations. Note also that in some cases the scheme is reformulated using time-varying constraints, e.g., in case of tube-based robust MPC, used in the safety-constrained MDP context in [28].

Finally, we remark that in [13] it has been proven that the value function, action-value function and policy can be obtained exactly by MPC, provided that the parameterization is rich enough. This argument can be proven to hold also in the case we consider in this paper, i..e, using undiscounted MPC with an inexact model to approximate a discounted MDP. A rigorous proof is omitted for the sake of brevity and because it is obtained following the same arguments used in [13].

5 Simulation Examples

In this section we provide two examples. The first one considers nominal asymptotic stability for a nonlinear system for which steady-state operation is not optimal. The second one considers a stochastic system for which we guarantee safety and stability by robust MPC.

5.1 Nominal Nonlinear Economic MPC

Consider the Continuously Stirred Tank Reactor (CSTR) from [11,2], where an irreversible chemical reaction A → B takes place with rate $k_\mathrm{r}c_\mathrm{A}$, where $k_\mathrm{r} = 0.4$ l/(mol min) and c_A, c_B are the concentrations of A and B respectively. The process dynamics are

$$\dot{c}_\mathrm{A} = \frac{q}{V_\mathrm{R}}(c_\mathrm{Af} - c_\mathrm{A}) - k_\mathrm{r}c_\mathrm{A}, \qquad \dot{c}_\mathrm{B} = \frac{q}{V_\mathrm{R}}(c_\mathrm{Bf} - c_\mathrm{B}) + k_\mathrm{r}c_\mathrm{A},$$

where $c_\mathrm{Af} = 1$ mol/l, $c_\mathrm{Bf} = 0$ mol/l are the feed concentrations of A and B, V_R is the volume of the reactor. The flow $q \in [0, 20]$ l/min through the reactor is the control variable. The system is discretized using sampling time $t_\mathrm{s} = 1$ min. The discount factor is $\gamma = 0.9$ and the stage cost is $\ell(\boldsymbol{s}, \boldsymbol{a}) = 2qc_\mathrm{A} - 1.5q$. As observed in [2], even though $\boldsymbol{s}_\mathrm{s} = (0.5, 0.5)$, $\boldsymbol{a}_\mathrm{s} = 4$ is an economically optimal steady-state, this cost does not yield asymptotic stability to that steady state for $\gamma = 1$, as periodic operation yields a lower cost. Also with $\gamma = 0.9$ periodic operation does yield a lower cost than operating the system at the optimal steady-state. We observe that, to the best of our knowledge, a formal method to solve this problem with formal steady-state stability guarantees is currently not available, with the exception of the approach we propose in this paper.

We formulate a nominal MPC scheme of the form (39) using the simple quadratic stage and initial cost

$$\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{u}) = \left\| \begin{matrix} \boldsymbol{x} - \boldsymbol{s}_\mathrm{s} \\ \boldsymbol{u} - \boldsymbol{a}_\mathrm{s} \end{matrix} \right\|_H^2,$$
$$\lambda_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left\| \boldsymbol{x} - \boldsymbol{s}_\mathrm{s} \right\|_\Lambda^2 + \boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{s}_\mathrm{s}) + l,$$

with parameter vector $\boldsymbol{\theta} = (H, \Lambda, \boldsymbol{\lambda}, l)$.

In order to obtain simple conditions for asymptotic stability, we enforce a terminal point constraint $\boldsymbol{x}_N = \boldsymbol{s}_\mathrm{s}$, with a prediction horizon $N = 100$. Since we enforce a terminal point constraint, a sufficient condition for nominal asymptotic stability is $H \succ \epsilon I$, with $\epsilon > 0$ [24]. Similarly to the approach proposed in [32], we formulate the problem similarly to Q-learning, i.e., as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathbb{E}\left[\left(Q_\star^\gamma(\boldsymbol{s}, \boldsymbol{a}) - Q_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{a}) \right)^2 \right] \\ \text{s.t.} \quad & H + \Delta H \succeq \epsilon I. \end{aligned}$$

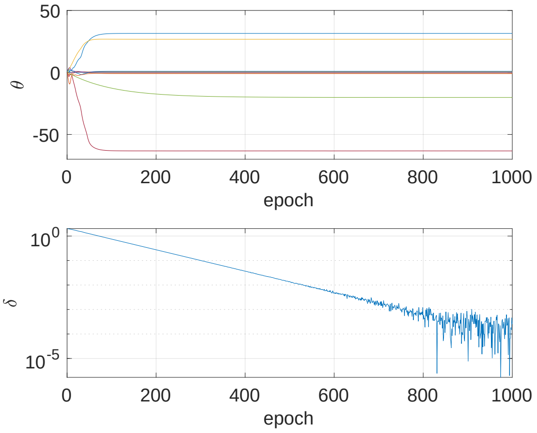

Fig. 2. Evolution of the parameter $\boldsymbol{\theta}$ and the TD error δ over the RL epochs.

We solve the problem by replacing $Q_\star^\gamma$ with its temporal-difference approximation in the optimality conditions. We select $\epsilon = 10^{-4}$ and, in each epoch, we collect a batch of 8 episodes of 40 samples each starting from initial conditions $(1, 0)$, $(0, 1)$, $(1, 1)$, $(0, 0)$, $(0.8, 0.3)$, $(0.3, 0.8)$, $(0.4, 0.6)$, $(0.6, 0.4)$. In each epoch, we take one single iterate of a quasi-Newton method based on the Gauss-Newton Hessian approximation. The obtained parameter update $\Delta\boldsymbol{\theta}$ is not applied directly, as we use the Q-learning like update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\Delta\boldsymbol{\theta}$ with learning rate $\alpha = 0.1$. Then we move to the next epoch, collect a new batch of data and iterate this procedure.

We ran a simulation over 1000 epochs and obtained the parameter evolution and TD-error displayed in Figure 2. We observe that the obtained stage cost matrix has eigenvalues 31.5, 10^{-4}, 10^{-4}, such that, as expected, the stability constraint is active.

5.2 Robust MPC

Consider the linear system with dynamics and stage cost

$$\boldsymbol{s}_+ = \begin{bmatrix} 1 & 0.1 \\ 0 & 1.1 \end{bmatrix} \boldsymbol{s} + \begin{bmatrix} 0.05 \\ 0.1 \end{bmatrix} \boldsymbol{a} + \boldsymbol{w},$$
$$\ell(\boldsymbol{s}, \boldsymbol{a}) = \begin{bmatrix} \boldsymbol{s} - \boldsymbol{s}^\mathrm{r} \\ \boldsymbol{a} - \boldsymbol{a}^\mathrm{r} \end{bmatrix}^\top \mathrm{diag}\left(\begin{bmatrix} 1 \\ 0.01 \\ 0.01 \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{s} - \boldsymbol{s}^\mathrm{r} \\ \boldsymbol{a} - \boldsymbol{a}^\mathrm{r} \end{bmatrix},$$

where $\boldsymbol{s} = (p, v)$ and $\boldsymbol{s}^{\mathrm{r}} = (-3, 0)$, $\boldsymbol{a}^{\mathrm{r}} = 0$, with discount factor $\gamma = 0.5$. Note that the discounted unconstrained problem without stability constraints is not stabilizing.

We formulate tube based MPC [9,30] as

$$\min_{\boldsymbol{z}} \sum_{k=0}^{N-1} \left\| \begin{matrix} \boldsymbol{x}_k - \boldsymbol{x}_{\mathrm{r}} \\ \boldsymbol{u}_k - \boldsymbol{u}_{\mathrm{r}} \end{matrix} \right\|_H^2 + \left\| \boldsymbol{x}_N - \boldsymbol{x}_{\mathrm{r}} \right\|_P^2 + \left\| \boldsymbol{x}_0 \right\|_\Lambda^2 + \boldsymbol{\lambda}^\top \boldsymbol{x}_0 + l \tag{40a}$$

$$\text{s.t. } \boldsymbol{x}_0 = \boldsymbol{s}, \qquad \boldsymbol{u}_0 = \boldsymbol{a}, \tag{40b}$$

$$\boldsymbol{x}_{k+1} = A\boldsymbol{x}_k + B\boldsymbol{u}_k + \boldsymbol{b}, \qquad k \in \mathbb{I}_0^{N-1}, \tag{40c}$$

$$C\boldsymbol{x}_k + D\boldsymbol{u}_k + \boldsymbol{c}_k \leq \boldsymbol{0}, \qquad k \in \mathbb{I}_0^{N-1}, \tag{40d}$$

$$T\boldsymbol{x}_N + \boldsymbol{t} \leq \boldsymbol{0}, \tag{40e}$$

where one must enforce that the system dynamics (40c) and a parameterized compact uncertainty set $\mathbb{W}_{\boldsymbol{\omega}}$ satisfy

$$\boldsymbol{s}_+ - (A\boldsymbol{s} + B\boldsymbol{a} + \boldsymbol{b}) \in \mathbb{W}_{\boldsymbol{\omega}}.$$

The polyhedral parameterization $\mathbb{W}_{\boldsymbol{\omega}} := \{\boldsymbol{w} | M\boldsymbol{w} \leq \boldsymbol{1}\}$ is used and the following set membership constraint is imposed on M for all observed samples $\boldsymbol{s}_{i+1}, \boldsymbol{s}_i, \boldsymbol{a}_i$, $i \in \mathcal{I}$:

$$M(\boldsymbol{s}_{i+1} - (A\boldsymbol{s}_i + B\boldsymbol{a}_i + \boldsymbol{b})) \leq \boldsymbol{1}, \qquad \forall\, i \in \mathcal{I}.$$

The constraints that the real system must satisfy are assumed to be given by $C\boldsymbol{s} + D\boldsymbol{a} + \hat{\boldsymbol{c}} \leq 0$. These constraints are tightened, i.e., $\boldsymbol{c}_k \geq \hat{\boldsymbol{c}}$, so as to guarantee that the original constraints are satisfied for any process noise $\boldsymbol{w} \in \mathbb{W}_{\boldsymbol{\omega}}$. Parameters $\boldsymbol{x}_{\mathrm{r}}, \boldsymbol{u}_{\mathrm{r}}$ must be a steady-state for the system dynamics (40c): $(A - I)\boldsymbol{x}_{\mathrm{r}} + B\boldsymbol{u}_{\mathrm{r}} = \boldsymbol{0}$. Finally, in order to have asymptotic stability to the minimum robust positively invariant set and recursive feasibility, T and $\boldsymbol{t}$ must be selected such that they define a robust positively invariant terminal set for the feedback law $\boldsymbol{u} = -K(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{r}}) + \boldsymbol{u}_{\mathrm{r}}$, with K the solution to the LQR formulated with A, B, H, P [9,30]. The vector of MPC parameters is then defined as

$$\boldsymbol{\theta} = \{\Lambda, \lambda, l, H, \boldsymbol{x}_{\mathrm{r}}, \boldsymbol{u}_{\mathrm{r}}, M\}, \tag{41}$$

while K, P, $\boldsymbol{c}_k$, T, $\boldsymbol{t}$ are functions of these parameters which guarantee that the terminal set is robust positively invariant by construction. Matrices C, D and vector $\hat{\boldsymbol{c}}$ are assumed to be known. Finally, A, B, $\boldsymbol{b}$ can in principle also be included in the parameter vector $\boldsymbol{\theta}$. However, as discussed in [28] modifying A and B makes the MDP much harder to formulate and solve.

For some small $\epsilon > 0$, the set of parameters guaranteeing safety and stability then becomes

$$\begin{aligned} \Theta := \{\, \boldsymbol{\theta} \mid\ & H \succeq \epsilon I, \\ & M(\boldsymbol{s}_{i+1} - (A\boldsymbol{s}_i + B\boldsymbol{a}_i + \boldsymbol{b})) \leq \boldsymbol{1},\ \forall\, i \in \mathcal{I}, \\ & (A - I)\boldsymbol{x}_{\mathrm{r}} + B\boldsymbol{u}_{\mathrm{r}} = \boldsymbol{0}, \\ & \exists\, \boldsymbol{x} \text{ s.t. } T(\boldsymbol{\theta})\boldsymbol{x} \leq \boldsymbol{t}(\boldsymbol{\theta}) \,\}, \end{aligned}$$

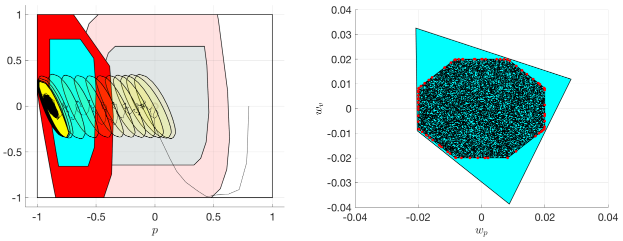

Fig. 3. Left plot: MRPI (red), terminal (cyan) sets and reference $\boldsymbol{x}^{\mathrm{r}}$ (black and gray circle) at the beginning (faded color) and end of the learning process (full color); state trajectory (black line) and mRPI sets (yellow) at each time instant. Right plot: true process noise set (transparent octagon), noise samples (black dots), their convex hull (red dots) and noise set parameterized by matrix M (cyan).

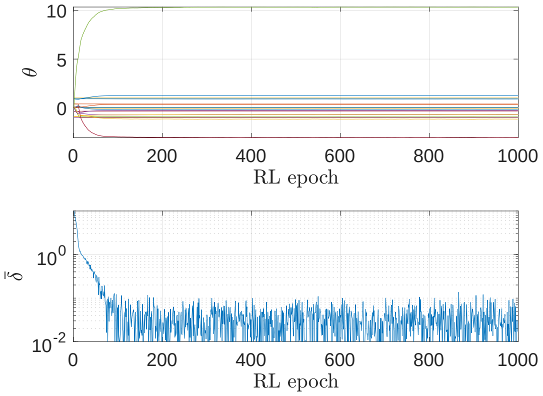

Fig. 4. Top plot: parameter evolution through the epochs. Bottom plot: TD error through the epochs.

i.e., the stage cost must be positive-definite, the noise set must include all observed noise samples, the reference must be a steady-state of the system and the terminal set must be nonempty. This last condition also entails that the MPC domain is nonempty.

We formulate tube based MPC (40) with prediction horizon $N = 50$ and define the state and control constraints as $-\boldsymbol{1} \leq \boldsymbol{s} \leq \boldsymbol{1}$, $-10 \leq \boldsymbol{a} \leq 10$. The real noise set is selected as a regular octagon, and we parameterize $\mathbb{W}_{\boldsymbol{\omega}}$ as a polytope with 4 facets. We update $\boldsymbol{\theta}$ using the constrained Q-learning approach proposed in [32] with learning factor $\alpha = 0.1$.

The closed-loop trajectory starting from $\boldsymbol{s}_0 = (0.8, 0)$ is displayed in Figure 3, together with the reference, Maximum Robust Positive Invariant (MRPI) and terminal sets at the beginning and end of the simulation, as well as the minimum Robust Positive Invariant (mRPI) sets throughout the simulation. We further display the noise set approximation at the end of the simulation in the same figure, and the evolution throughout the RL epochs of the parameter $\boldsymbol{\theta}$ and the average TD error in each batch in Figure 4.

6 Conclusions

In this paper, we have provided a way to enforce stability constraints in discounted MDPs. To that end, we have proved that, under a weak assumption, any discounted MDP can be reformulated as an undiscounted MDP, and we have proved that the obtained undiscounted MDP yields bias optimal policies. In order to introduce the stability constraints, we exploited the results of [13] to deploy stabilizing MPC formulations to support the value functions and policy approximations required to solve the MDP. Future work will further investigate the stability properties of discounted MDPs and the possibility of using the equivalence in order to derive new algorithms for solving undiscounted MDPs.

References

[1] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *In Advances in Neural Information Processing Systems 19*, page 2007. MIT Press, 2007.

[2] R. Amrit, J. Rawlings, and D. Angeli. Economic optimization using model predictive control with a terminal cost. *Annual Reviews in Control*, 35:178–186, 2011.

[3] A. Aswani, H. Gonzalez, S.S. Sastry, and C. Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216 – 1226, 2013.

[4] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe Model-based Reinforcement Learning with Stability Guarantees. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 908–918. Curran Associates, Inc., 2017.

[5] D. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 3rd edition, 2005.

[6] D. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, 3rd edition, 2007.

[7] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

[8] D. Blackwell. Discrete Dynamic Programming. *The Annals of Mathematical Statistics*, pages 719–726, 1962.

[9] L. Chisci, J.A. Rossiter, and G. Zappa. Systems with persistent disturbances: predictive control with restricted constraints. *Automatica*, 37:1019–1028, 2001.

[10] S. Dean, S. Tu, N. Matni, and B. Recht. Safely Learning to Control the Constrained Linear Quadratic Regulator. In *2019 American Control Conference (ACC)*, pages 5582–5588, July 2019.

[11] M. Diehl, R. Amrit, and J.B. Rawlings. A Lyapunov Function for Economic Optimizing Model Predictive Control. *IEEE Trans. of Automatic Control*, 56(3):703–707, March 2011.

[12] V. Gaitsgory, L. Grüne, M. Höger, C. M. Kellett, and S. R. Weller. Stabilization of strictly dissipative discrete time systems with discounted optimal control. *Automatica*, 93:311 – 320, 2018.

[13] S. Gros and M. Zanon. Data-Driven Economic NMPC Using Reinforcement Learning. *IEEE Transactions on Automatic Control*, 65(2):636–648, Feb 2020.

[14] S. Gros and M. Zanon. Reinforcement Learning for Mixed-Integer Problems Based on MPC. In *21st IFAC World Congress*, volume 53, pages 5219–5224, 2020.

[15] S. Gros and M. Zanon. An Economic MPC Dissipativity Theory for Undiscounted Markov Decision Processes. *Automatica*, 2022. (submitted).

[16] S. Gros, M. Zanon, and A. Bemporad. Safe Reinforcement Learning via Projection on a Safe Set: How to Achieve Optimality? In *21st IFAC World Congress*, volume 53, pages 8076–8081, 2020.

[17] L. Grüne and J. Pannek. *Nonlinear Model Predictive Control*. Springer, London, 2011.

[18] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause. Learning-based Model Predictive Control for Safe Exploration and Reinforcement Learning. Published on Arxiv, 2018.

[19] D.Q. Mayne, M.M. Seron, and S.V. Rakovic. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41:219–224, 2005.

[20] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, USA, 2nd edition, 2009.

[21] Chris J. Ostafew, Angela P. Schoellig, and Timothy D. Barfoot. Robust Constrained Learning-based NMPC enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016.

[22] R. Postoyan, L. Buşoniu, D. Nešić', and J. Daafouz. Stability of infinite-horizon optimal control with discounted cost. In *53rd IEEE Conference on Decision and Control*, pages 3903–3908, 2014.

[23] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[24] J. B. Rawlings, D. Q. Mayne, and M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2 edition, 2017.

[25] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, USA, 2nd edition, 2018.

[26] K. Wabersich, L. Hewing, A. Carron, and M. Zeilinger. Probabilistic model predictive safety certification for learning-based control. *arXiv:1906.10417v1, 25 Jun 2019*, 2019.

[27] X. Wang, J. Stoev, G. Pinte, and J. Swevers. Energy optimal point-to-point motion using Model Predictive Control. In *Proceedings ASME 2012 5th Annual Dynamic Systems and Control Conference*, 2012.

[28] M. Zanon and Gros. Safe Reinforcement Learning Using Robust MPC. *IEEE Transactions on Automatic Control*, 66(8):3638–3652, 2021.

[29] M. Zanon and S. Gros. Reinforcement Learning Based on Real-Time Iteration NMPC. In *IFAC World Congress*, 2020.

[30] M. Zanon and S. Gros. On the Similarity Between Two Popular Tube MPC Formulations. In *Proceedings of the European Control Conference*, pages 651–656, 2021.

[31] M. Zanon and S. Gros. A New Dissipativity Condition for Asymptotic Stability of Discounted Economic MPC. *Automatica*, 2022. (in press).

[32] M. Zanon, S. Gros, and A. Bemporad. Practical Reinforcement Learning of Stabilizing Economic MPC. In *Proceedings of the European Control Conference*, pages 2258–2263, 2019.